KNOWLEDGE DISCOVERY WITH HYBRID DATA MINING APPROACH

A
SYNOPSIS

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE

BY
DEEPAK SARASWAT

UNDER THE SUPERVISION OF
Prof. Preetvanti Singh
FACULTY OF SCIENCE

DEPARTMENT OF PHYSICS AND COMPUTER SCIENCE
FACULTY OF SCIENCE
DAYALBAGH EDUCATIONAL INSTITUTE
DAYALBAGH, AGRA

September 2016

HEAD
Department of Physics and Computer Science
Faculty of Science

DEAN
Faculty of Science

INTRODUCTION

Due to tremendous advancement in digital data collection devices, computing power, and data storage technology there is an explosive growth in the amount of stored data. This data contains valuable hidden knowledge which could be used to improve the decision-making process. Thus there is an urgent need for a new generation of (semi-)automatic methods for extracting knowledge from the data. These tools and techniques are the subject of the emerging field of knowledge discovery in databases (Fayyad, Piatetsky-Shapiro & Smyth, 1996a).

Knowledge Discovery

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. The focus of KDD is finding understandable patterns that can be interpreted as useful knowledge. It can be defined as the organized process of identifying valid, novel, potentially useful, and understandable patterns from large and complex data sets (Fayyad, Piatetsky-Shapiro & Smyth, 1996b). The three general properties that the discovered knowledge should satisfy are accuracy, comprehensibility, and interestingness:

- A decision maker may often be interested in discovering knowledge which has a certain predictive power, i.e. predicting the values for future based on previously observed data. Thus, the discovered knowledge must have a high predictive accuracy rate.

- The discovered knowledge should be comprehensible for supporting effective decision making, i.e. the discovered knowledge should simply not be a black box that makes predictions without explaining it.

- A pattern is interesting if it is easily understood, novel, valid on new or test data with some degree of certainty, potentially useful, or validates some hypothesis that a user seeks to confirm. Subjective interestingness measures may be based on user's belief in the data, for example, unexpectedness, novelty, and action ability. On the other hand objective interestingness measures are usually based on statistics and structures of patterns, for example, support and confidence.

The Knowledge Discovery Process

A KDD framework provides an overview of major knowledge discovery activities, and connects them using a process flow which has raw data as input and usable knowledge as the output (Fayyad, Piatetsky-Shapiro & Smyth, 1996a). The knowledge discovery process is inherently iterative as shown in Figure 1. The output of a step can not only be sent to next step in the process, but can also be sent as a feedback to the previous steps. This process includes the application of several pre-processing and post-processing methods aimed at refining and improving the discovered knowledge. The KDD process includes the following steps:

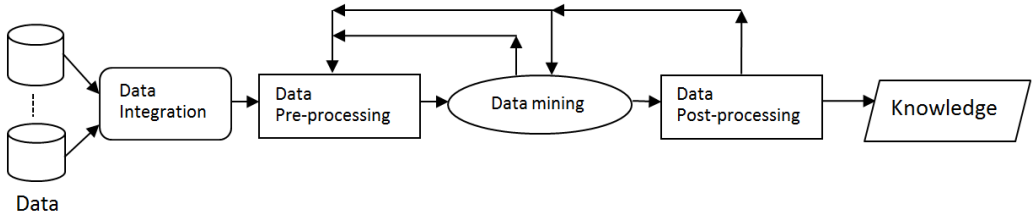

Figure 1: An overview of the knowledge discovery process

(a) *Data Selection*- This step first requires learning the application domain by analyzing relevant prior knowledge and goals of the applications. Next the target data set is created.

Two types of goals can be identified for KDD (based on the intended use of the system):

- ✓ **Verification** which includes verifying the user's hypothesis, and
- ✓ **Discovery** which means autonomously finding new patterns. Discovery goal can further be divided into
  - **Prediction,** where the system finds patterns for predicting the future behavior of some entities; and
  - **Description**, where the system finds patterns for presenting them to a user in a human-understandable form.

(b) *Data Integration* - This is necessary if the data comes from several different sources.

(c) *Data Cleaning* - It is important to make sure that the data is as accurate as possible. This step involves detecting and correcting errors in the data, or filling in missing values.

(d) *Discretization* - This step consists of transforming a continuous attribute into a categorical (or nominal) attribute, taking on only a few discrete values. Discretization often improves the comprehensibility of the discovered knowledge (Peng, Kou, Shi & Chen, 2008).

(e) *Data mining*- This step includes selecting a specific method for searching pattern in a representable form.

(f) *Interpretation*- Interpreting the discovered pattern for decision making.

Data Mining

Data Mining is the core of the knowledge discovery process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. Data mining is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology to extract high level knowledge from real-world data sets.

Data mining involves determining pattern from or fitting models to observed data. It is sophisticated data analytical method that focuses upon exploration and devoloping new insights for supporting decision making. This extracted information is useful for identifying trends, forming a prediction or classification model and summarizing a database.

Data mining Techniques

The data mining techniques include:

1. *Classification/Regression:* discovery of a model or function that maps objects into predefined classes (classification) or into suitable values (regression). The model/function is computed on a training set (supervised learning).

2. *Clustering:* identifying a finite set of categories or clusters to describe the data.

3. *Summarization:* finding a compact description for a subset of data, for example the derivation of summary or association rules and the use of multivariate visualization techniques.

4. *Dependency Modeling:* finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.

5. *Sequential Patterns:* discovery of frequent subsequences in a collection of sequences (sequence database), each representing a set of events occurring at subsequent times. The ordering of the events in the subsequences is relevant.

6. *Change and Deviation Detection:* discovering the most significant changes in the data from previously measured or normative values.

Taxonomy of Data Mining Methods
As discussed above there are many methods of Data Mining which are used for different purposes and goals. The two main types of Data Mining are: verification-oriented (the system verifies the user's hypothesis) and discovery-oriented (the system finds new rules and patterns autonomously).

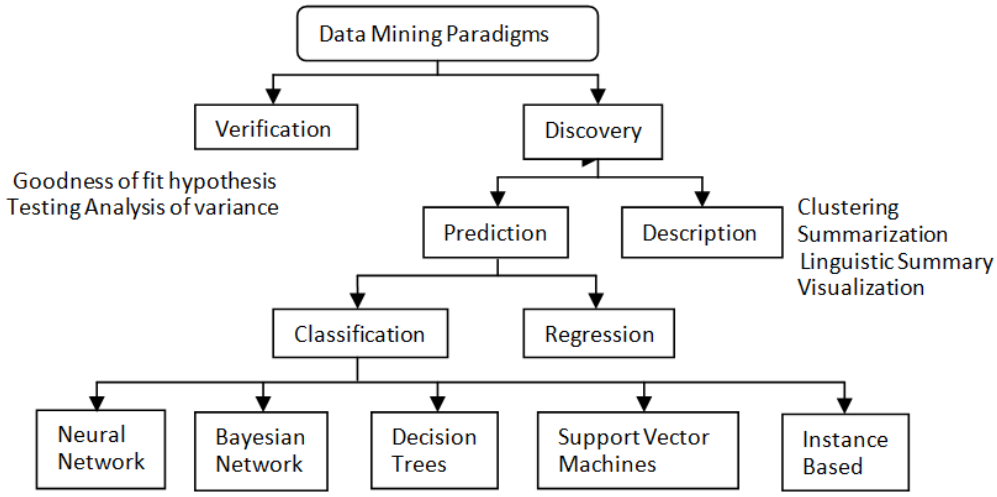

Figure 2: Taxonomy of Data Mining

Discovery methods are those that automatically identify patterns in the data. The discovery method branch consists of prediction and description methods. Descriptive methods are oriented to data interpretation, which focuses on understanding the way the underlying data relates to its parts. Prediction-oriented methods aim to build a behavioral model, which obtains new and unseen samples and is able to predict values of one or more variables related to the sample. It also develops patterns which form the discovered knowledge in a way which is understandable and easy to operate upon.

Verification methods, on the other hand, deal with the evaluation of a hypothesis proposed by an external source (like an expert). These methods include traditional statistical methods like goodness of fit test, tests of hypotheses, and analysis of variance (ANOVA).

The Components of Data Mining Algorithms
The three primary components in any data mining algorithm are: *model representation, model evaluation,* and *search.*

a. *Model Representation:* This is the language used to describe discoverable patterns. It is important that a data analyst fully comprehend the representational assumptions which may be inherent in a particular method.

b. *Model Evaluation Criteria:* These are the quantitative statements (or fit functions) of how well a particular pattern (a model and its parameters) meet the goals of the KDD process. Descriptive models can be evaluated along the dimensions of predictive accuracy, novelty, utility, and understandability of the fitted model.

c. *Search Method:* consists of two components: *parameter search* and *model search.* Once the model representation and the model evaluation criteria are fixed, then the data mining problem is reduced to

an optimization task, i.e. finding the parameters/models from the selected family which optimizes the evaluation criteria.

In *parameter search* the algorithm must search for the parameters which optimize the model evaluation criteria given observed data and a fixed model representation. *Model search* occurs as a loop over the parameter search method, the model representation is changed so that the family of models is considered.

Nowadays, decision makers invariably need to use decision support technology in order to tackle with complex decision making problems. In this area, data mining has an important role to extract valuable information. Consequently, the use of data mining and decision support methods, including novel visualization methods, can lead to better performance in decision making, and enable tackling of new types of problems that have not been addressed before.

LITERATURE REVIEW

Knowledge provides power in many real-life contexts enabling and facilitating the preservation of valuable heritage, new learning, solving intricate problems, creating core competencies and initiating new situations for both individuals and organizations now and in the future (Choudhary, Harding & Lin, 2007). The huge amounts of data in databases, which contain large numbers of records, with many attributes that need to be simultaneously explored to discover useful information and knowledge, make manual analysis impractical. All these factors indicate the need for intelligent and automated data analysis methodologies, which might discover useful knowledge from data. Knowledge discovery in databases and data mining have therefore become extremely important tools in realizing the objective of intelligent and automated data analysis. Data mining is a carefully planned application of statistical and machine-learning methods and tools through the space of analytic techniques which is considered as a process of deciding what will be most useful, promising, and revealing. A detailed review of data mining tools and their applications can be found in Deshpande & Thakare (2010), and Ramageri (2010). Broadly, the major tasks of data mining are predictive and descriptive tasks under discovery oriented data mining system.

PREDICTIVE TASKS

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The tasks under predictive modeling include classification, prediction and time-series analysis.

Classification

Classification is finding models that analyze and classify a data item into several predefined classes. Common tools used for classification are decision trees, neural network, Bayesian network and rough set theory. Friedman, Wolff & Schuster (2008) presented extended definitions of k-anonymity and used them to prove that a given data mining model does not violate the k-anonymity of the individuals represented in the learning examples. The authors demonstrated that the developed model can be applied to various data mining problems, such as classification, association rule mining and clustering. Liu, Sharma & Datla (2008) discussed the adaptability of available imputation techniques for holiday traffic and then introduced a new procedure using non-parametric regression, the k-nearest neighbor (k-NN) method. A subspace-based multimedia data mining framework was proposed by Shyu, Xie, Chen & Chen (2008) for video semantic analysis, specifically video event/concept detection, by addressing two basic issues, i.e., semantic gap and rare event/concept detection. Siradeghyan, Zakarian & Mohanty (2008) presented a new associative classification algorithm for data mining. The algorithm uses elementary set concepts, information entropy and database manipulation techniques to develop useful relationships between input and output attributes of large databases. These relationships (knowledge) are represented using IF–THEN association rules. Yeh (2008) applied genetic algorithms for selecting a group of relevant genes from cancer microarray data. Then, the popular classifiers, such as OneR, Naïve-Bayes, decision tree, and Support Vector Machine (SVM) were built on the basis of these selected genes. A novel approach to the problem of detecting and classifying underwater bottom mine objects in littoral environments from acoustic backscattered signals was considered by Wang (2009). The author defined robust short-time Fourier transform to convert the received echo into a time–frequency plane. To identify the decision parameters that affect the feasibility analysis, data mining techniques were applied by Yun & Caldas (2009) to analyze the Go/No Go decision-making process in infrastructure projects. Zhang, Kinsner & Huang (2009) presented an electrocardiogram (ECG) data mining scheme based on the ECG frame classification realized by a Dynamic Time Warping matching technique.

A novel architecture was proposed by Frantzidis et al. (2010) for the robust discrimination of emotional physiological signals evoked upon viewing pictures selected from the International Affective Picture System. The biosignals were initially differentiated according to their valence dimension by means of a data mining approach, the C4.5 decision tree algorithm. Shi (2010) proposed and extended a series of

optimization-based classification models via multiple criteria linear programming and multiple criteria quadratic programming. Wang & Wang (2010) presented the conceptual foundations of data mining with incomplete data through classification which is relevant to a specific decision making problem. The proposed technique assumes incomplete data and complete data may come from different sub-populations. Pecchia, Melillo & Bracale (2011) proposed a platform to enhance effectiveness and efficiency of home monitoring using data mining for early detection of any worsening in patient's condition. The authors designed the remote health monitoring platform which supports heart failure severity assessment offering functions of data mining based on classification and regression tree method. Grossi & Turini (2012) outlined novel data structures and algorithms when the model mined out of the data is a classifier. The introduced model and the overall ensemble architecture was presented in details, considering how the approach can be extended for treating numerical attributes. The shape prior segmentation procedure and pruned association rule with Image Apriori algorithm was used by Rajendran & Madheswaran (2012) to develop an improved brain image classification system. The CT scan brain images have been classified into three categories namely normal, benign and malignant, considering the low-level features extracted from the images and high level knowledge from specialists to enhance the accuracy in decision process.

Dogan & Tanrikulu (2013) compared the classification algorithm accuracy, speed (CPU time consumed) and robustness for various datasets and their implementation techniques. The data miner selects the model mainly with respect to classification accuracy. Erdogan (2013) applied SVMs to bank bankruptcy analysis. Which were implemented to analyze financial ratios. Data sets from Turkish commercial banks were used. Mohanty, Senapati & Lenka (2013) developed a computer-aided classification system for cancer detection from digital mammograms. The proposed system consists of three major steps. The first step is region of interest extraction of 256*256 pixels size. The second step is the feature extraction. The third step is the classification process, where the technique of the association rule mining was used to classify between normal and cancerous tissues. Mokhtar & Elsayad (2013) analyzed data mining classification algorithms; Decision Tree (DT), Artificial Neural Network (ANN), and SVM on mammographic masses dataset to increase the ability of physicians in determining the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. Jeyakumar, Li & Suthaharan (2014) studied SVM classifiers in the face of uncertain knowledge sets and showed how data uncertainty in knowledge sets can be treated in SVM classification by employing robust optimization. For distributed data mining in peer-to-peer systems Kokkinos & Margaritis (2014) described a completely asynchronous, scalable and privacy-preserving committee machine. Regularization neural networks were used for all the peer classifiers.

Different data mining methods: classification and regression trees, random forest and M5 rules, were tested by Ruiz-Samblas, Cadenas, Pelta & Cuadros-Rodríguez (2014) for classification and prediction. These approaches were also used for feature selection prior to modeling in order to reduce the number of attributes in the chromatogram. In order to cover the shortage of the classical association rules optimized algorithm. Zheng & Wang (2014) proposed an optimized method of frequent sets calculation, a method of parallel NEclat combined with cloud programming model. Lin, Chang & Chen (2015) proposed a scheme for privacy-preserving outsourcing. The service provider solved the SVM from the perturbed data without knowing the actual content of the data. The generated SVM classifier was also perturbed, which would only be recovered by the data owner. Maxwell, Warner, Strager, Conley & Sharp (2015) investigated machine-learning algorithms and measures derived from RapidEye satellite imagery, and light detection and ranging (lidar) data for geographic object-based image analysis classification of mining and mine reclamation. SVMs, random forests, and boosted classification and regression trees classification algorithms were assessed and compared with the k-nearest neighbor (k-NN) classifier.

Ram & Doegar (2015) used data mining techniques DT and ANN for the classification of Statlog heart disease datasets. These supervise machine learning algorithms were compared on the basis of

classification accuracy and performance matrices. Castro & Kim (2016) used three data mining classification models to detect factors with the greatest influence on car accidents. The experimental objective explored the role of different factors on injury risk using a Bayesian network, DT and ANN.
Karan & Samadder (2016) evaluated the performance of SVM-based image classification technique with the maximum likelihood classification technique for a rapidly changing landscape of an open-cast mine. The authors also assessed the change in land use pattern due to coal mining from 2006 to 2016. Tang, He, Baggenstoss & Kay (2016) introduced a Bayesian classification approach for automatic text categorization by using class-specific features. Unlike conventional text categorization approaches, Authors proposed a method to selects a specific feature subset for each class. For applying these class-specific features for classification, Authors followed Baggenstoss's PDF Projection Theorem (PPT) to reconstruct the PDFs in raw data space from the class-specific PDFs in low-dimensional feature subspace, and build a Bayesian classification rule. Taylor et al. (2016) presented a data mining methodology for driving condition monitoring via controller area network-bus data that is based on the general data mining process. The approach is applicable to many driving condition problems, and the example of road type classification without the use of location information is investigated.

Prediction

Bellazzi & Zupan (2008) provided a comprehensive review of predictive data mining in clinical medicine and gave guidelines to carry out data mining studies in this field. Sun & Li (2008) developed a data mining method combining attribute-oriented induction, information gain, and decision tree, which is suitable for preprocessing financial data and constructing decision tree model for financial distress prediction. Models for short- and long-term prediction of wind farm power were discussed by Kusiak, Zheng & Song (2009). The models were built using weather forecasting data generated at different time scales and horizons. The wind farm power prediction models were built with five different data mining algorithms. Weber & Mateas (2009) presented a data mining approach to opponent modeling in strategy games. The authors also discussed how to incorporate the developed data mining approach into a full game playing agent. A survey cum experimental methodology was adopted by Ramaswami & Bhaskaran (2010) to generate a database from a primary and a secondary source. The raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection. A set of prediction rules was extracted from CHIAD prediction model. Srinivas, Rani & Govrdhan (2010) briefly examined the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. Vu & Khan (2010) reported a model for real-time prediction of urban bus running time based on statistical pattern recognition technique, the locally weighted scatter smoothing. The trained model automatically searches through the historical patterns which are the most similar to the current pattern and on that basis, the prediction is made. Baradwaj & Pal (2011) used the classification task to evaluate student's performance using the decision tree method. Authors extracted knowledge that describes students' performance in end semester examination that helps earlier in identifying the dropouts and students who need special attention. Kumari & Godara (2011) analyzed data mining classification techniques RIPPER classifier, DT, ANN, and SVM on cardiovascular disease dataset. Authors used 10-fold cross validation method to measure the unbiased estimate of these prediction models Maroco et al. (2011) advanced the hypothesis that newer statistical classification methods derived from data mining and machine learning methods. Model predictors were 10 neuropsychological tests currently used in the diagnosis of dementia. Statistical distributions of classification parameters obtained from a 5-fold cross-validation were compared using the Friedman's nonparametric test. Bhatla & Jyoti (2012) analyzed various data mining techniques introduced in recent years for heart disease prediction. The observations reveal that neural networks with 15 attributes has outperformed over all other data mining techniques. Dangare & Apte (2012) analysed prediction systems for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease.

The authors added two more attributes i.e. obesity and smoking. Kotsiantis (2012) presented a case study for predicting students' marks. Students' key demographic characteristics and their marks in a small number of written assignments can constitute the training set for a regression method in order to predict the student's performance. Olson, Delen & Meng (2012) applied a variety of data mining tools to bankruptcy data, with the purpose of comparing accuracy and number of rules. For this data, decision trees were found to be relatively more accurate compared to neural networks and SVMs, but there were more rule nodes than desired. Verbeke, Dejaeger, Martens, Hur & Baesens (2012) developed a profit centric approach to evaluate customer churn prediction models and reported a wide benchmarking study on data mining for churn prediction. Amin, Agarwal & Beg (2013) presented a technique for prediction of heart disease using major risk factors. This technique involved two most successful data mining tools, neural networks and genetic algorithms. The hybrid system implemented uses the global optimization advantage of genetic algorithm for initialization of neural network weights. Dogra & Wala (2015) studied classification and clustering techniques on the basis of algorithms which were used to predict previously unknown class of objects and presented a detailed description of data mining techniques and algorithms. Dimitriado, Papaemmanouil & Diao (2016) developed AIDE, an Automatic Interactive Data Exploration framework that assists users in discovering new interesting data patterns and eliminate expensive ad-hoc exploratory queries. AIDE relies on a seamless integration of classification algorithms and data management optimization techniques that collectively strive to accurately learn the user interests based on his relevance feedback on strategically collected samples. AIDE can deliver highly accurate query predictions for very common conjunctive queries with small user effort while, given a reasonable number of samples, it can predict with high accuracy complex disjunctive queries.

Dindarloo & Siami-Irdemoosa (2016) explored the application of classification and clustering approaches for pattern recognition and failure forecasting on mining shovels. The shovels were classified into four clusters using k-means clustering algorithms. Future failures were predicted using the SVM classification technique.

Time-series Analysis

A time series is a collection of observations made chronologically. The nature of time series data includes: large in data size, high dimensionality and necessary to update continuously. The increasing use of time series data has initiated a great deal of research and development attempts in the field of data mining. Batyrshin & Sheremetov (2008) considered architecture of perception-based decision making system in time series databases domains integrating perception-based TSDM, computing with words and perceptions, and expert knowledge. The new tasks which should be solved by the perception-based TSDM methods to enable their integration in such systems are also discussed. In order to provide a comprehensive validation, Ding, Trajcevski, Scheuermann, Wang & Keogh (2008) conducted an extensive set of time series experiments re-implementing 8 different representation methods and 9 similarity measures and their variants, and testing their effectiveness on 38 time series data sets from a wide variety of application domains. Tang, Yang & Zhou (2009) combined news mining and time series analysis to forecast inter-day stock prices. News reports are automatically analyzed with text mining techniques, and then the mining results are used to improve the accuracy of time series analysis algorithms. Ye & Keogh (2009) introduced a new time series primitive, the time series shapelets and demonstrated with extensive empirical evaluations in diverse domains. Zubcoff, Pardillo &Trujillo (2009) proposed a unified modelling language (UML) extension through UML profiles for data-mining. The extension provides analysts with an intuitive notation for time-series analysis which is independent of any specific data-mining tool or algorithm. A comprehensive revision on the existing time series data mining research is given by Fu (2011). The primary objective was to serve as a glossary for interested researchers to have an overall picture on the current time series data mining development. Chen, Hong & Tseng (2012) extended previous fuzzy mining approach for handling time-series data to find linguistic association rules. The proposed approach first uses a sliding window to generate continues subsequences from a given time series and then analyzes the fuzzy itemsets from these subsequences. Appropriate post-

processing is then performed to remove redundant patterns. Rakthanmanon et al. (2012) showed that by using a combination of four novel ideas massive time series can truly be searched and mined truly. Authors demonstrated their work on the largest set of time series experiments ever attempted. Vieira et al. (2012) developed a methodology for contributing in the automation of sugarcane mapping over large areas, with time-series of remotely sensed imagery. Two major techniques were combined: Object Based Image Analysis (OBIA) and Data Mining (DM). OBIA was used to represent the knowledge needed to map sugarcane, whereas DM was applied to generate the knowledge model. Lee & Kam (2014) explored and discussed the potential use of time-series data mining, a relatively new framework by integrating conventional time-series analysis and data mining techniques, to discover actionable insights and knowledge from the transportation temporal data. A case study on the Singapore public train transit was used to demonstrate the time-series data-mining framework and methodology.

DESCRIPTIVE TASKS

The objective here is to derive patterns that summarize the underlying relationship in data. The tasks under descriptive modeling are clustering, summarizing and dependency modeling.

Clustering

Clustering is identifying a finite set of categories or clusters to describe the data. Common tools used for clustering include k-means, principal component analysis, the Kolmogorov-Smirnov test, the quantile range test, and polar ordination. Hu, Meng & Shi (2008) proposed a new validity function of fuzzy clustering for spatial data. Mrázová & Dagli (2008) proposed a novel approach to assign an adequate semantics to clusters formed by fuzzy c-means clustering. The introduced fuzzy c-landmarks show a great potential for dimension reduction and for simplified data set descriptions. Wang, Zhang, Wang & Lai (2008) presented a kernel clustering-based SVM (KCB-SVM) that generalizes the linear clustering-based SVM (CB-SVM) to solve nonlinear classification problems in a novel way. Liao, Chen & Hsu (2009) applied Apriori algorithm, and performed clustering analysis based on an ontology-based data mining approach for mining customer knowledge from the database. Knowledge extracted from data mining results is illustrated as knowledge patterns, rules, and maps in order to propose suggestions and solutions to the case firm for possible product promotion and sport marketing. Mehar, Maeder, Matawie, & Ginige (2010) suggested an approach which can offer greater range of choice for generating potential clusters of interest. An example on health services utilization characterization according to socio-demographic background is discussed and the blended clustering approach being taken for it is described. Zamani, Pourmand & Saree (2010) applied hierarchical cluster analysis to aid in the development of traffic signal timing plans. This approach can be used for designing of a time-of-day (TOD) signal control system, since it automatically identifies TOD intervals using the historical collected data. Gecchele, Rossi, Gastaldi & Caprini (2011) presented a comparative analysis of various data mining clustering methods for the grouping of roads, aimed at the estimation of Annual Average Daily Traffic. Lee & Estivill-Castro (2011) incorporated two data mining techniques, clustering and association-rule mining, into a exploratory tool for the discovery of spatial-temporal patterns in data-rich environments. This tool is an autonomous pattern detector that efficiently and effectively reveals plausible cause–effect associations among many geographical layers. Density based clustering algorithm is one of the primary methods for clustering in data mining. The clusters which are formed based on the density are easy to understand and it does not limit itself to the shapes of clusters. Parimala, Lopez and Senthilkumar (2011) gave a detailed survey of the existing density based algorithms.

Kumar (2012) applied Fuzzy K-Means clustering on healthcare data set to reduce the formal context and FCA on the reduced data set for mining association rules. Wan, Lei & Chou (2012) developed a Landslide Expert System by using multi-date SPOT image data to develop the landslide database. The threshold slope which becomes vulnerable to landslides is obtained by the K-means method. Fan, Bouguila & Ziou (2013) proposed a variational inference framework for unsupervised non- Gaussian

feature selection, in the context of finite generalized Dirichlet (GD) mixture-based clustering. Under the proposed principled variational framework, authors simultaneously estimated, in a closed form, all the involved parameters and determined the complexity (i.e., both model a feature selection) of the GD mixture.

Ghosh & Lohani (2013) carried out an assessment of available categories of clustering techniques and found that hierarchical- and density based algorithms are apt for clustering light detection and ranging (lidar) data. The authors also adapted and examined the effect of two algorithms, density-based spatial clustering of applications with noise (DBSCAN) and ordering of points to identify the clustering structure (based on perimeter of triangles) (OPTICS (BOPT)) in the area of knowledge discovery in databases, on lidar data. Subspace clustering finds sets of objects that are homogeneous in subspaces of high-dimensional datasets. Sim, Gopalkrishnan, Zimek & Cong (2013) presented the basic subspace clustering and the related works in high-dimensional clustering. A clustering algorithm with shape control was introduced by Tabesh & Askari-Nasab (2013) which can provide reasonable guidelines for all the shapes by calibrating its parameters. The implementations of the algorithm on two small datasets with 874 and 2794 blocks were also illustrated.

Mozafary & Payvandy (2014) applied data-mining technique in textile industry by using data-mining methods containing clustering and ANN. The results showed that the performance of data-mining technique is more accurate than that of ANN. A novel fuzzy based unsupervised clustering algorithm proposed by Thomas & Raju (2014) was extended to segment quantitative values into fuzzy clusters. Membership values of quantitative items in the partitioning fuzzy clusters were used with weighted fuzzy rule mining techniques to find natural association rules. Wang & Eick (2014) proposed a polygon-based clustering and analysis framework for mining multiple geospatial datasets that have inherently hidden relations. Hung, Peng & Lee (2015) proposed a new trajectory pattern mining framework, the Clustering and Aggregating Clues of Trajectories (CACT), for discovering trajectory routes that represent the frequent movement behaviors of a user. Zime & Vreeken (2015) introduced the fundamental problems of different sub-fields of clustering, especially focusing on subspace clustering, ensemble clustering, alternative (as a variant of constraint) clustering, and multiview clustering (as a variant of alternative clustering). Then the authors related a representative of subspace clustering to pattern mining.

Gabor filtering, image post processing, feature construction through application of principal component analysis, k-means clustering and first level classification using Naïve–Bayes classification algorithm and second level classification using C4.5 enhanced with bagging techniques were applied by Geetharamani & Balasubramanian (2015) to help ophthalmologists in providing early treatment to the patients. Kumar & Toshniwal (2015) proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of 11,574 road accidents on road network of Dehradun (India) between 2009 and 2014. Next, association rule mining were used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set and the clusters identified by K-modes clustering algorithm. Kumar & Toshniwal (2016) applied k-means algorithm to group the accident locations into three categories, high-frequency, moderate-frequency and low-frequency accident locations. k-means algorithm takes accident frequency count as a parameter to cluster the locations. Then association rule mining was used to characterize these locations. The rules revealed different factors associated with road accidents at different locations with varying accident frequencies. Zhang, Zhang, Liu & Liu (2016) introduced a multi-task multi-view clustering framework which integrates within-view-task clustering, multi-view relationship learning, and multi-task relationship learning. Under this framework, authors proposed two multi-task multi-view clustering algorithms, the bipartite graph based multi-task multi-view clustering algorithm, and the semi-nonnegative matrix tri-factorization based multi-task multi-view clustering algorithm. The former one can deal with the multi-task multi-view clustering of nonnegative data; the latter one is a general multi-task multi-view clustering method.

Summarization

Summarization is finding a compact description for a subset of data. The tools for summarization are association rule mining and optimization. Zambreno, Özıs. Ikyılmaz, Memik & Choudhary (2006) presented a set of representative data mining applications called MineBench. The authors evaluate the MineBench applications on an 8-way shared memory machine and analyzed some important performance characteristics. A multi-knowledge based approach was proposed by Zhuang, Jing & Zhu (2006) which integrates WordNet, statistical analysis and movie knowledge. The experimental results show the effectiveness of the proposed approach in movie review mining and summarization. Chandola & Kumar (2007) formulated the problem of summarization of a dataset of transactions with categorical attributes as an optimization problem involving two objective functions - compaction gain and information loss. Authors proposed metrics to characterize the output of any summarization algorithm. Özıs, Ikyılmaz, Narayanan, Zambreno, Memik et al. (2006) presented MineBench, a publicly available benchmark suite containing fifteen representative data mining applications belonging to various categories: classification, clustering, association rule mining and optimization. Haghighi et al. (2013) presented a novel generic toolkit that enables building situation and resource-aware mobile data mining applications and describe along with underlying theoretical foundations of resource and situation criticality, awareness and adaptation, which are entirely transparent and hidden from the user.

Dependency Modeling

This is concerned with finding a model that describes significant relationships between attribute sets. Common tools for dependency modeling are *Apriori* association rules and sequential pattern analysis. Hwang, Chang, Chen, & Wu (2008) used data mining in healthcare to develop clinical pathway guidelines and provided an evidence-based medicine platform. It could give better results in knowledge refinement through a use of this technique on the construction industry dataset (Ur-Rahman, & Harding, 2012). Liao, Chen & Wu (2008) used to mine customer knowledge from household customers. A multidimensional mining approach is presented by Behnisch & Ultsch (2009) as a case study applied to 12 430 German communities to analyze multidynamic characteristics between 1994 and 2004. In particular, Emergent Self Organizing Maps was performed as an appropriate method for clustering and classification. Denton & Wu (2009) considered sets of related continuous attributes as vector data and searched for patterns that relate a vector attribute to one or more items.

Duan, Street & Xu (2011) used correlations among nursing diagnoses, outcomes and interventions to create a recommender system for constructing nursing care plans. The system utilizes a prefix-tree structure common in itemset mining to construct a ranked list of suggested care plan items based on previously-entered items. Iakovidis & Smailis (2012) presented a novel semantic model that describes knowledge extracted from the lowest-level of a data mining process, where information is represented by multiple features i.e. measurements or numerical descriptors extracted from measurements, images, texts or other medical data, forming multidimensional feature spaces. This model enables a unified representation of knowledge across multimodal data repositories. Kim, Chae & Olson (2013) identified three sets of direct marketing data with a different degree of class imbalance (little, moderate, high) and used random under sampling method to reduce the degree of the imbalance problem. Ko, Hong, Choi & Kim (2013) developed a wafer-to-wafer fault detection system using data stream mining techniques for a semiconductor etch tool. The system consists of two data stream mining modules: a trace segmentation module and a multivariate trace comparison module. Ofoghi, Zeleznikow, MacMahon & Raab (2013) investigated different data mining demands of elite sports with respect to a number of features that describe sport competitions. The aim is to more structurally connect the sports and data mining domains through: (a) describing a framework for categorizing elite sports, and (b) understanding the analytical demands of different performance analysis problems. Günnemann, Färber, Boden & Seid (2014) proposed a new method to find homogeneous object groups in a single vertex-labeled graph.

Panov, Soldatova & Dzeroski (2014) presented Onto-core, an ontology of core data mining entities. Onto-core defines the most essential data mining entities in a three-layered ontological structure comprising of a specification, an implementation and an application layer. Sim, Choi & Kim (2014) developed a data mining approach to which large amounts of trace data are inputted to infer fault-introducing machines in the form of a L⇒R rule, where R contains the fault type and L contains a machine sequence that is the primary cause of the fault type. Mirge, Verma & Gupta (2016) presented a novel technique for excavating heavy traffic flow patterns in bi-directional road network, i.e. identifying divisions of the roads where the traffic flow is very dense. The proposed technique works in two phases: phase I finds the clusters of trajectory points based on density of trajectory points; phase II arranges the clusters in sequence based on spatiotemporal values for each route and directions.

Motivation of proposed work

From the above discussion following are the main observations:

- The computerization of many business and government transactions and the advances in data collection tools provide huge amount of data. This explosive growth has generated an urgent need for new data mining tools and techniques that can intelligently and automatically transform the processed data into useful information and knowledge (Fayyad, Piatetsky-Shapiro & Smyth, 1996b).

  One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets. Researchers from database systems, artificial intelligence, machine learning, knowledge acquisition, statistics, spatial databases, and information providing services have shown great interests in data mining techniques. This means that data mining has great importance from the real-life application viewpoint in advancement of the service provided and in increasing the business opportunities.

- Data mining is receiving more and more attention from the business community, as witnessed by frequent publications in the popular IT-press, and the growing number of tools appearing on the market (Feelders, Daniels, & Holsheimer, 2000).

  Data Mining in Society: Data Mining is primarily used in every field to "drill down" data and determine data pattern like customer preferences, product positioning, impact on sales and so on.
  - ✓ Data mining holds great potential to improve health systems. Researchers used data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. (Koh and Tan, (2011); Chaurasia and Pal (2014); Sudhakar and Manimekalai (2014); Roski, Bo-Linn and Andrews (2014)).
  - ✓ Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational context. It could be oriented towards students in order to recommend learners' activities, and improve learning skills (Liao and Liao (2011); Natek and Zwilling (2014); Baker (2014);Xing, Petakovic and Goggins (2015)).
  - ✓ Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data (Lin and Harding (2007); Köksal, Batmaz and Testik (2011);Ur-Rahman and Harding (2012); Wu, Wang and Schaefer (2015)).
  - ✓ Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because of the volume of data. The techniques may be used to find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer(Chen and Du (2009); Aburrous, Hossain, Dahaland Thabtah (2010); Koh, Tan and Goh (2015); Geng, Bose and Chen(2015)).
  - ✓ Mobile phone and utilities companies use Data Mining and Business Intelligence to predict 'churn', the terms they use for when a customer leaves their company to get their

phone/gas/broadband from another provider. (Chen, Chiang and Storey (2012); Kaplan (2012); Hung, Yen and Wang (2006); Lazer, Kennedy, King and Vespignani (2014)).

Because the business environment is so dynamic, it is often difficult for businesses to quickly identify emerging patterns or trends. Data Mining Tools help businesses identify problems and opportunities promptly and then make quick and appropriate decisions with the new business intelligence which can be used to improve vital business processes.

- The traditional use of data mining through its software tools does not bring it closer to business users due to the complexity of data mining tools, thus there is a need to collaborate data mining and decision support system (Khademolqorani & Hamadani, 2013; Chen, 2016).

  The use of data mining to facilitate decision support enables new approaches to problem solving by discovering patterns and relationships hidden in data and therefore enabling an inductive approach to decision support system.

It is also observed single data mining technique is not useful for gathering the appropiate information from multi-databases therefore data mining must be integrated with some multi-criteria decision making or decision support system for solving realistic intelligent buisness problems.

PROPOSED WORK

This research work is devoted to develop method for knowledge discovery with hybrid data mining approach to provide much useful and powerful decision.

Objectives of the proposed study

1. To develop hybrid data mining method for multi-databases.
2. To incorporate data mining component in the decision support system.
3. To present case study that shows applicability of the proposed intelligent miner to a real-life application.

Variables

This study will deal with the variables based on the application area chosen for the study area. Many practical data mining systems divide attributes into two types: categorical corresponding to nominal, binary and ordinal variables; and continuous corresponding to integer, interval-scaled and ratio-scaled variables. A third category of attribute, the 'ignore' attribute, may also be considered corresponding to variables that are of no significance for the application but which cannot be deleted from the dataset.

Methodology

1. **Development of the hybrid data mining method:** This research study will be devoted to the study of discovery-oriented data mining method. Techniques like multi-criteria decision making and/or case based reasoning will be integrated with a data mining technique for identifying patterns from multi-databases (Object-relational database or Spatial and temporal data, the one best suited in this study will be taken).

2. **Integrating hybrid data mining and decision support system:** A decision support system will be designed and developed to incorporate the hybrid data mining model and its solution procedure preferably using MATLAB/.Net Framework.

3. **Selecting and creating a dataset on which discovery will be performed**
   Having defined the goals, the data that will be used for the knowledge discovery will be determined. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process. From the literature review, it was observed that most of data mining techniques are applied on medical/ healthcare problems. The application domain for this study will be any one: e-commerce, or sports depending upon the applicability of the developed method.

Data Mining within a Decision Support System

Today a decision maker invariably uses decision support technology to tackle with a complex decision making problem. Decision Support System allows a user to intuitively, quickly, and flexibly manipulate data to provide analytical insight. Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends.

The use of data mining methods and decision support methods can lead to better performance in decision making. Figure 3 shows the DSS process diagram through described methods:

1. **Understanding of the application domain:** This step requires understanding the business objectives and requirements, and then converting this knowledge into a decision making problem definition and a preliminary plan design to achieve the objectives. Also, one should get familiar with the data to

discover first insights into the data or to detect interesting subsets to form hypotheses about hidden information.

2. **Data preprocessing**- As part of a data pre-processing-step, Dimensional reduction is extremely important in real-world applications as it has been effective in removing duplicates, increasing learning accuracy, and improving decision making processes.

   Among linear and non-linear sampling, and similarity measure dimensionality reduction the process best suited in this research study will be applied.
3. **Data Integration**: These data should be integrated with available database.

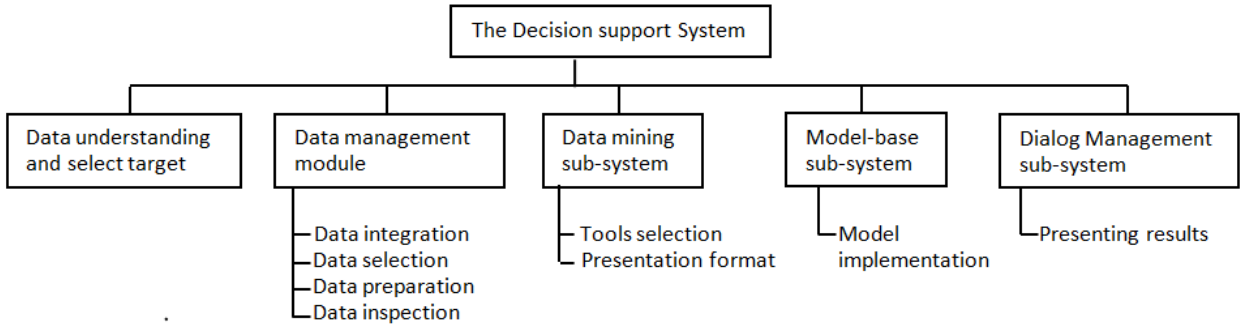

Figure 3: The decision support process

4. **Data selection**: The next step is a subset of integrated database that should be classified.
5. **Data preparation**: Data preparation includes all required activities for constructing the final data set, i.e. the data that will be fed into the modeling tool.
6. **Data inspection**: At the last step of this module, data inspection is implied to evaluate prepared data for analysis.

The *Database Management Sub-System* of the proposed decision support system will be responsible for data integration, Data Selection, Data Preparation and Data Inspection tasks.

7. **Tools selection:** Data Mining uses methods, algorithms, and techniques from a variety of disciplines to extract useful knowledge from large amounts of data in order to support decision making. Data mining Tool Selection can be based on understanding the data and selecting the target selection together with data inspection.

8. **Presentation format:** This step requires developing the formats of outputs to be utilized by managers and end users.

The *Data Mining Sub-System* of the proposed decision support system will be responsible for Tools Selection, and Presentation Format.

The methods will also extend the possibilities of interpreting the data, and discovering information, trends and patterns by using richer model representations (e.g. decision rules, frequent patterns)

9. **Model Implementation:** Finally, Model implementation starts data mining process which contains classification, clustering, prediction, and /or association rules mining. In addition to these usual tasks, the decision makers can use the multi-criteria decision making methods to rank and prioritize the group of options, and optimize multi-objectives.

The *Model-base Management Sub-System* of the proposed decision support system will be responsible for Model Implementation.

Dialog Management Sub-System

The Dialog Management Sub-System is one of the main sub-systems of the decision support system as it fills the space between data miners and business users.

The Semantic protocol diagram is shown in Figure 4.

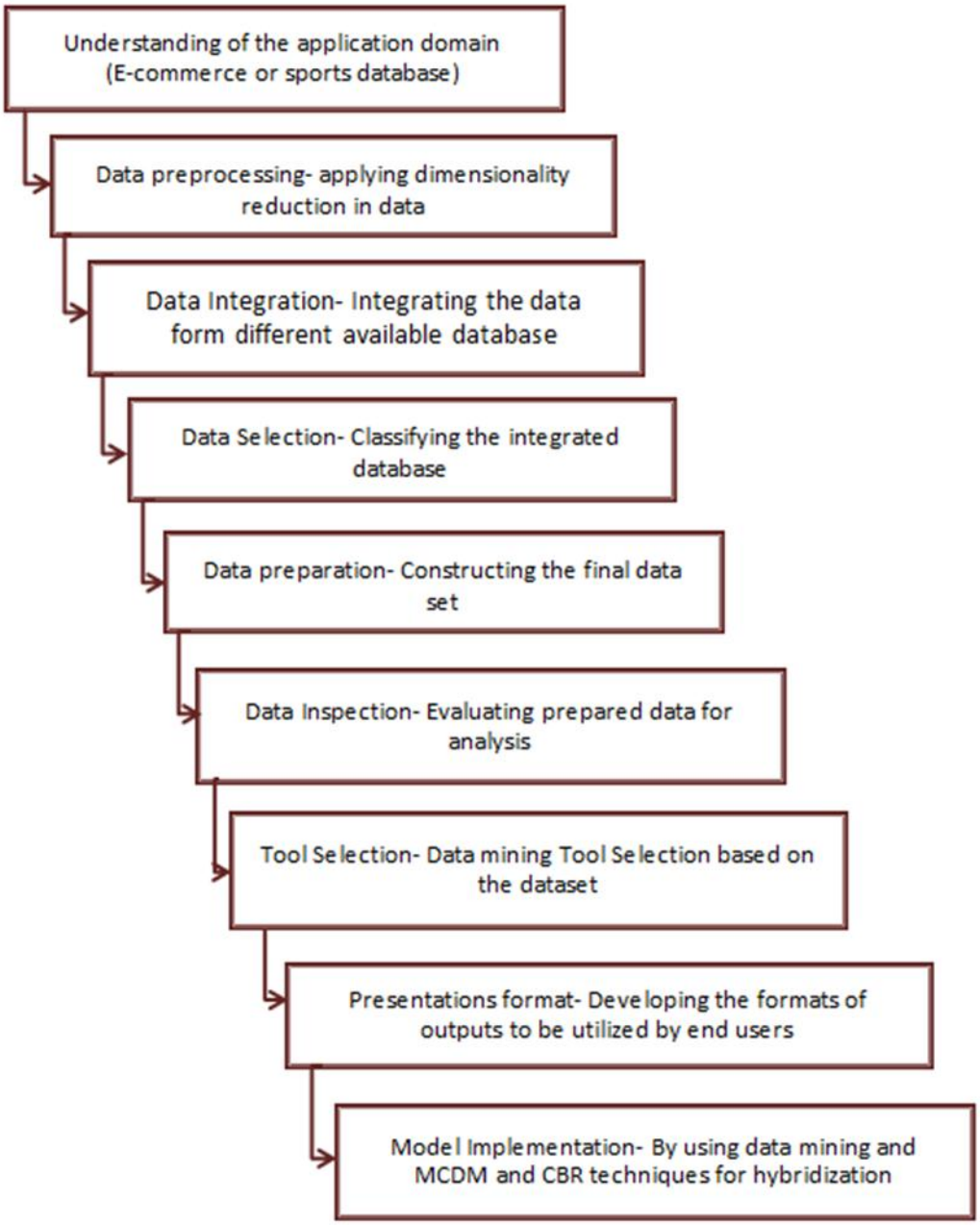
Understanding of the application domain (E-commerce or sports database)
Data preprocessing- applying dimensionality reduction in data
Data Integration- Integrating the data form different available database
Data Selection- Classifying the integrated database
Data preparation- Constructing the final data set
Data Inspection- Evaluating prepared data for analysis
Tool Selection- Data mining Tool Selection based on the dataset
Presentations format- Developing the formats of outputs to be utilized by end users
Model Implementation- By using data mining and MCDM and CBR techniques for hybridization

Figure 4: Semantic protocol

It is expected that the developed intelligent miner will provide much useful, reliable and powerful decision, and can be useful for analysts to gain business intelligence by identifying and observing trends, problems and anomalies.

References

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert systems with applications, 37*(12), 7913-7921.

Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus, 4*(1), 1.

Amin, S. U., Agarwal, K., & Beg, R. (2013). Genetic neural network based data mining in prediction of heart disease using risk factors. In *IEEE Conference on Information & Communication Technologies (ICT), 2013*. 1227-1231.

Baker, R. S. (2014). Educational Data Mining: An Advance for Intelligent Systems in Education. *IEEE Intelligent Systems, 29*(3), 78-82.

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *(IJACSA) International Journal of Advanced Computer Science and Applications, 2*(6), 63-69.

Batyrshin, I. Z., & Sheremetov, L. B. (2008). Perception-based approach to time series data mining. *Applied Soft Computing, 8*(3), 1211-1221.

Behnisch, M., & Ultsch, A. (2009). Urban data-mining: spatiotemporal exploration of multidimensional data. *Building Research & Information, 37*(5-6), 520-532.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics, 77*(2), 81-97.

Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering, 1*(8), 1-4.

Castro, Y., & Kim, Y. J. (2016). Data mining on road safety: factor assessment on vehicle accidents using classification models. *International Journal of Crashworthiness, 21*(2), 104-111.

Chandola, V., & Kumar, V. (2007). Summarization–compressing data into an informative representation. *Knowledge and Information Systems, 12*(3), 355-378.

Chaurasia, V., & Pal, S. (2014). Data Mining Approach to Detect Heart Diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2*, 56-66.

Chen, C. H., Hong, T. P., & Tseng, V. S. (2012). Fuzzy data mining for time-series data. *Applied Soft Computing, 12*(1), 536-542.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly, 36*(4), 1165-1188.

Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus, 5*(1), 1-16.

Chen, W. S., & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert systems with applications, 36*(2), 4075-4086.

Choudhary, A., Harding, J., & Lin, H. (2007). Engineering moderator to universal knowledge moderator for moderating collaborative projects. *Global Journal of e-Business & Knowledge Management, 3*(1), 5-12.

Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications, 47*(10), 44-48.

De Montjoye, Y. A., Radaelli, L., & Singh, V. K. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science, 347*(6221), 536-539.

Denton, A. M., & Wu, J. (2009). Data mining of vector–item patterns using neighborhood histograms. *Knowledge and Information Systems, 21*(2), 173-199.

Deshpande, S. P., & Thakare, V. M. (2010). Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS), 1*(1), 32-44.

Dimitriadou, K., Papaemmanouil, O., & Diao, Y. (2016). AIDE: An Active Learning-Based Approach for Interactive Data Exploration. *IEEE Transactions on Knowledge and Data Engineering, 28*(11), 2842-2856.

Dindarloo, S. R., & Siami-Irdemoosa, E. (2015). Data mining in mining engineering: results of classification and clustering of shovels failures data. *International Journal of Mining, Reclamation and Environment*, 1-14.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment, 1*(2), 1542-1552.

Dogan, N., & Tanrikulu, Z. (2013).A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. *Information Technology and Management, 14*(2), 105-124.

Dogra, A. K., & Wala, T. (2015). A review paper on data mining techniques and algorithms. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), 4*(5), 1976-1979.

Duan, L., Street, W. N., & Xu, E. (2011). Healthcare information systems: data mining methods in the creation of a clinical recommender system. *Enterprise Information Systems, 5*(2), 169-181.

Erdogan, B. E. (2013). Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. *Journal of Statistical Computation and Simulation, 83*(8), 1543-1555.

Fan, W., Bouguila, N., & Ziou, D. (2013). Unsupervised hybrid feature extraction selection for high-dimensional non-Gaussian data clustering with variational inference. *IEEE Transactions on Knowledge and Data Engineering, 25*(7), 1670-1685.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996a). Knowledge discovery and data mining: towards a unifying framework. *In KDD* (*Vol. 96*, pp. 82-88).

Fayyd, U. M., Shapiro, G. P., & Smyth, P. (1996b). From data mining to knowledge discovery: an overview.

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information & Management, 37*(5), 271-281.

Frantzidis, C. A., Bratsas, C., Klados, M. A., Konstantinidis, E., Lithari, C. D., Vivas, A. B., ... & Bamidis, P. D. (2010). On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine, 14*(2), 309-318.

Friedman, A., Wolff, R., & Schuster, A. (2008). Providing k-anonymity in data mining. *The VLDB Journal, 17*(4), 789-804.

Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence, 24*(1), 164-181.

Gecchele, G., Rossi, R., Gastaldi, M., & Caprini, A. (2011). Data mining methods for traffic monitoring data analysis: A case study. *Procedia-Social and Behavioral Sciences, 20*, 455-464.

Geetharamani, R., & Balasubramanian, L. (2015). Automatic segmentation of blood vessels from retinal fundus images through image processing and data mining techniques. *Sadhana, 40*(6), 1715-1736.

Ghosh, S., & Lohani, B. (2013). Mining lidar data with spatial clustering algorithms. *International Journal of Remote Sensing, 34*(14), 5119-5135.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236-247.

Grossi, V., & Turini, F. (2012). Stream mining: a novel architecture for ensemble-based classification. *Knowledge and Information Systems, 30*(2), 247-281.

Günnemann, S., Färber, I., Boden, B., & Seidl, T. (2014).GAMer: a synthesis of subspace clustering and dense subgraph mining. *Knowledge and Information Systems, 40*(2), 243-278.

Haghighi, P. D., Krishnaswamy, S., Zaslavsky, A., Gaber, M. M., Sinha, A., & Gillick, B. (2013). Open mobile miner: A toolkit for building situation-aware data mining applications. *Journal of Organizational Computing and Electronic Commerce, 23*(3), 224-248.

Hu, C., Meng, L., & Shi, W. (2008).Fuzzy clustering validity for spatial data. *Geo-spatial Information Science, 11*(3), 191-196.

Hung, C. C., Peng, W. C., & Lee, W. C. (2015). Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. *The VLDB Journal, 24*(2), 169-192.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(3), 515-524.

Hwang, H. G., Chang, I. C., Chen, F. J., & Wu, S. Y. (2008). Investigation of the application of KMS for diseases classifications: A study in a Taiwanese hospital. *Expert Systems with Applications, 34*(1), 725-733.

Iakovidis, D., & Smailis, C. (2012). A semantic model for multimodal data mining in healthcare information systems. *Stud Health Technol Inform, 180*, 574-578.

Jeyakumar, V., Li, G., & Suthaharan, S. (2014). Support vector machine classifiers with uncertain knowledge sets via robust optimization. *Optimization, 63*(7), 1099-1116.

Jurca, G., Addam, O., Aksac, A., Gao, S., Özyer, T., Demetrick, D., & Alhajj, R. (2016). Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. *BMC Research Notes, 9*(1), 1.

Kaplan, A. M. (2012). If you love something, let it go mobile: Mobile marketing and mobile social media 4x4. *Business horizons, 55*(2), 129-139.

Karan, S. K., & Samadder, S. R. (2016). Accuracy of land use change detection using support vector machine and maximum likelihood techniques for open-cast coal mining areas. *Environmental Monitoring and Assessment, 188*(8), 486.

Khademolqorani, S., & Hamadani, A. Z. (2013). An adjusted decision support system through data mining and multiple criteria decision making. *Procedia-Social and Behavioral Sciences, 73*, 388-395.

Kim, G., Chae, B. K., & Olson, D. L. (2013). A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models. *Service Business, 7*(1), 167-182.

Ko, J. M., Hong, S. R., Choi, J. Y., & Kim, C. O. (2013). Wafer-to-wafer process fault detection using data stream mining techniques. *International Journal of Precision Engineering and Manufacturing, 14*(1), 103-113.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management, 19*(2), 65.

Koh, H. C., Tan, W. C., & Goh, C. P. (2015). A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information, 1*(1).

Kokkinos, Y., & Margaritis, K. G. (2014).A distributed privacy-preserving regularization network committee machine of isolated Peer classifiers for P2P data mining. *Artificial Intelligence Review, 42*(3), 385-402.

Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert systems with Applications, 38*(10), 13448-13467.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review, 37*(4), 331-344.

Kumar, C. A. (2012). Fuzzy clustering-based formal concept analysis for association rules mining. *Applied Artificial Intelligence, 26*(3), 274-301.

Kumar, S., & Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data, 2*(1), 1.

Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation, 24*(1), 62-72.

Kumari, M., & Godara, S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction. *International Journal of Computer Science and Technology, 2*(2), 304-308.

Kusiak, A., Zheng, H., & Song, Z. (2009). Wind farm power prediction: a data-mining approach. *Wind Energy, 12*(3), 275-293.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science, 343*(6176), 1203-1205.

Lee, I., & Estivill-Castro, V. (2011). Exploration of massive crime data sets through data mining techniques. *Applied Artificial Intelligence, 25*(5), 362-379.

Lee, R. K. W., & Kam, T. S. (2014). Time-series data mining in transportation: a case study on singapore public train commuter travel patterns. *International Journal of Engineering and Technology, 6*(5), 431-438.

Liao, S. H., Chen, C. M., & Wu, C. H. (2008). Mining customer knowledge for product line and brand extension in retailing. *Expert Systems with Applications, 34*(3), 1763-1776.

Liao, S. H., Chen, J. L., & Hsu, T. Y. (2009). Ontology-based data mining approach implemented for sport marketing. *Expert Systems with Applications, 36*(8), 11045-11056.

Lin, H. K., & Harding, J. A. (2007). A manufacturing system engineering ontology model on the semantic web for inter-enterprise collaboration. *Computers in Industry, 58*(5), 428-437.

Lin, K. P., Chang, Y. W., & Chen, M. S. (2015). Secure support vector machines outsourcing with random linear transformation. *Knowledge and Information Systems, 44*(1), 147-176.

Liu, Z., Sharma, S., & Datla, S. (2008). Imputation of missing traffic data during holiday periods. *Transportation Planning and Technology, 31*(5), 525-544.

Luhang, X. (2015).The research of data mining in traffic flow data. *International Journal of Database Theory and Application, 8*(4), 19-30.

Manjunath, T. N., Hegadi, R. S., Umesh, I. M., &Ravikumar, G. K. (2012).Realistic analysis of data warehousing and data mining application in education domain. *International Journal of Machine Learning and Computing, 2*(4), 419.

Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M., & Bakker, P. J. (2008). Application of process mining in healthcare–a case study in a dutch hospital. *In International Joint Conference on Biomedical Engineering Systems and Technologies*. 425-438.

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes, 4*(1), 299.

Maxwell, A. E., Warner, T. A., Strager, M. P., Conley, J. F., & Sharp, A. L. (2015).Assessing machine-learning algorithms and image-and lidar-derived variables for GEOBIA classification of mining and mine reclamation. *International Journal of Remote Sensing, 36*(4), 954-978.

Mehar, A. M., Maeder, A., Matawie, K., & Ginige, A. (2010). Blended clustering for health data mining. In *E-Health* (pp. 130-137). *Springer Berlin Heidelberg*.

Mirge, V., Verma, K., & Gupta, S. (2016). Dense traffic flow patterns mining in bi-directional road networks using density based trajectory clustering. *Advances in Data Analysis and Classification*, 1-15.

Mohanty, A. K., Senapati, M. R., & Lenka, S. K. (2013). An improved data mining technique for classification and detection of breast cancer from mammograms. *Neural Computing and Applications, 22*(1), 303-310.

Mokhtar, S. A., & Elsayad, A. (2013).Predicting the severity of breast masses with data mining methods. *IJCSI International Journal of Computer Science Issues, 10*(2), 160-168.

Mozafary, V., & Payvandy, P. (2014). Application of data mining technique in predicting worsted spun yarn quality. *The Journal of The Textile Institute, 105*(1), 100-108.

Mrazova, I., & Dagli, C. H. (2008). Semantic clustering of the World Bank data¥ This research was partially supported by the grant No. 201/04/2102 of the GA ČR, by the Program "Information Society" under project 1ET100300517 and by the Grant Agency of Charles University in Prague under Grant No. 358/2006/A-INF/MFF. *International Journal of General Systems, 37*(4), 417-442.

Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert systems with applications, 41*(14), 6400-6407.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data mining in elite sports: a review and a framework. *Measurement in Physical Education and Exercise Science, 17*(3), 171-186.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, 52*(2), 464-473.

Ozisikyilmaz, B., Narayanan, R., Zambreno, J., Memik, G., & Choudhary, A. (2006, October). An architectural characterization study of data mining and bioinformatics workloads. In *2006 IEEE International Symposium on Workload Characterization* (pp. 61-70). IEEE.

Panov, P., Soldatova, L., & Džeroski, S. (2014). Ontology of core data mining entities. *Data Mining and Knowledge Discovery, 28*(5-6), 1222-1265.

Parimala, M., Lopez, D., & Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology, 31*(1), 59-66.

Pecchia, L., Melillo, P., & Bracale, M. (2011). Remote health monitoring of heart failure with data mining via CART method on HRV features. *IEEE Transactions on Biomedical Engineering, 58*(3), 800-804.

Peng, Y., Kou, G., Shi, Y., & Chen, Z. (2008). A descriptive framework for the field of data mining and knowledge discovery. *International Journal of Information Technology & Decision Making, 7*(04), 639-682.

Rajendran, P., & Madheswaran, M. (2012). An improved brain image classification technique with mining and shape prior segmentation procedure. *Journal of Medical Systems, 36*(2), 747-764.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., ... & Keogh, E. (2012, August). Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 262-270). ACM.

Ram, S., & Doegar, A. (2015). A Comparative Study of Data Mining Techniques for Predicting Disease Using Statlog Heart Disease Database. *International Journal of Advanced Research in Computer Science and Software Engineering, 5*(6), 5.

Ramageri, B. M. (2010). Data mining techniques, and applications. *Indian Journal of Computer Science and Engineering, 1*(4), 301-305.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *IJCSI International Journal of Computer Science Issues, 7*(1), 10-18.

Roski, J., Bo-Linn, G. W., & Andrews, T. A. (2014). Creating value in health care through big data: opportunities and policy implications. *Health Affairs, 33*(7), 1115-1122.

Ruiz-Samblás, C., Cadenas, J. M., Pelta, D. A., & Cuadros-Rodríguez, L. (2014). Application of data mining methods for classification and prediction of olive oil blends with other vegetable oils. *Analytical and Bioanalytical Chemistry, 406*(11), 2591-2601.

Shi, Y. (2010). Multiple criteria optimization-based data mining methods and applications: a systematic survey. *Knowledge and Information Systems, 24*(3), 369-391.

Shyu, M. L., Xie, Z., Chen, M., & Chen, S. C. (2008). Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transactions on Multimedia, 10*(2), 252-259.

Sim, H., Choi, D., & Kim, C. O. (2014). A data mining approach to the causal analysis of product faults in multi-stage PCB manufacturing. *International Journal of Precision Engineering and Manufacturing, 15*(8), 1563-1573.

Sim, K., Gopalkrishnan, V., Zimek, A., & Cong, G. (2013). A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery, 26*(2), 332-397.

Siradeghyan, Y., Zakarian, A., & Mohanty, P. (2008). Entropy-based associative classification algorithm for mining manufacturing data. *International Journal of Computer Integrated Manufacturing, 21*(7), 825-838.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE), 2*(02), 250-255.

Sudhakar, K., & Manimekalai, D. M. (2014). Study of heart disease prediction using data mining. *International journal of advanced research in computer science and software engineering, 4*(1).

Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems, 21*(1), 1-5.

Tabesh, M., & Askari-Nasab, H. (2013). Automatic creation of mining polygons using hierarchical clustering techniques. *Journal of Mining Science, 49*(3), 426-440.

Tang, B., He, H., Baggenstoss, P. M., & Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering, 28*(6), 1602-1606.

Tang, X., Yang, C., & Zhou, J. (2009). Stock price forecasting by combining news mining and time series analysis. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09.* 279-282.

Taylor, P., Griffiths, N., Bhalerao, A., Anand, S., Popham, T., Xu, Z., & Gelencser, A. (2016). Data mining for vehicle telemetry.*Applied Artificial Intelligence, 30*(3), 233-256.

Thomas, B. & Raju, G. (2014). A novel unsupervised fuzzy clustering method for preprocessing of quantitative attributes in association rule mining. *Information Technology and Management, 15*, 9-17.

Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications, 39*(5), 4729-4739.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211-229.

Vieira, M. A., Formaggio, A. R., Rennó, C. D., Atzberger, C., Aguiar, D. A., & Mello, M. P. (2012). Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. *Remote Sensing of Environment, 123*, 553-562.

Vu, N. H., & Khan, A. M. (2010).Bus running time prediction using a statistical pattern recognition technique. *Transportation Planning and Technology, 33*(7), 625-642.

Wan, S., Lei, T. C., & Chou, T. Y. (2012). A landslide expert system: image classification through integration of data mining approaches for multi-category analysis. *International Journal of Geographical Information Science, 26*(4), 747-770.

Wang, H., & Wang, S. (2010). Mining incomplete survey data through classification. *Knowledge and Information Systems, 24*(2), 221-233.

Wang, Q. (2009). Underwater bottom still mine classification using robust time–frequency feature and relevance vector machine. *International Journal of Computer Mathematics, 86*(5), 794-806.

Wang, S., & Eick, C. F. (2014).A polygon-based clustering and analysis framework for mining spatial datasets. *GeoInformatica, 18*(3), 569-594.

Wang, Y. H., & Liao, H. C. (2011). Data mining for adaptive learning in a TESL-based e-learning system. *Expert Systems with Applications, 38*(6), 6480-6485.

Wang, Y., Zhang, X., Wang, S., & Lai, K. K. (2008).Nonlinear clustering-based support vector machine for large data sets. *Optimization Methods & Software, 23*(4), 533-549.

Weber, B. G., & Mateas, M. (2009, September). A data mining approach to strategy prediction. In *2009 IEEE Symposium on Computational Intelligence and Games*. 140-147.

Wu, D., Rosen, D. W., Wang, L., & Schaefer, D. (2015). Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation. *Computer-Aided Design, 59*, 1-14.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47*, 168-181.

Ye, L., & Keogh, E. (2009, June). Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 947-956.

Yeh, J. Y. (2008). Applying data mining techniques for cancer classification on gene expression data. *Cybernetics and Systems: An International Journal, 39*(6), 583-602.

Yun, S., & Caldas, C. H. (2009).Analysing decision variables that influence preliminary feasibility studies using data mining techniques. *Construction Management and Economics, 27*(1), 73-87.

Zamani, Z., Pourmand, M., & Saraee, M. H. (2010). Application of data mining in traffic management: case of city of Isfahan. *In International Conference on Electronic Computer Technology (ICECT)*, 102-106.

Zambreno, J., Ikyılmaz, B. Ö., Memik, G., & Choudhary, A. (2006). Performance characterization of data mining applications using MineBench. In *In 9th Workshop on Computer Architecture Evaluation using Commercial Workloads (CAECW).*

Zhang, G., Kinsner, W., & Huang, B. (2009). Electrocardiogram data mining based on frame classification by dynamic time warping matching. *Computer Methods in Biomechanics and Biomedical Engineering, 12*(6), 701-707.

Zhang, X., Zhang, X., Liu, H., & Liu, X. (2016). Multi-Task Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering, 28*(12), 3324-3338.

Zheng, X., & Wang, S. (2014). Study on the method of road transport management information data mining based on pruning eclat algorithm and map reduce. *Procedia-Social and Behavioral Sciences, 138*, 757-766.

Zhuang, L., Jing, F., & Zhu, X. Y. (2006, November). Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 43-50). ACM.

Zimek, A., &Vreeken, J. (2015). The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Machine Learning, 98*(1-2), 121-155.

Zubcoff, J., Pardillo, J., & Trujillo, J. (2009). A UML profile for the conceptual modelling of data-mining with time-series in data warehouses. *Information and Software Technology, 51*(6), 977-992.